融合宽残差和长短时记忆网络的动态手势识别研究 *

梁智杰 [1,2]，廖盛斌 [1]

(1. 华中师范大学 国家数字化学习工程技术研究中心，武汉 430079; 2. 西南科技大学 网络教育学院，四川 绵阳 621010)

摘　要：针对现有的动态手势识别方法对长时间序列的时空特征难以精确匹配的问题，提出了一种基于宽残差和双向长短时记忆网络的时空特征一致手势识别方法。首先使用已经训练好的 3D 卷积神经网络从视频的空间和时间维度同步提取出短时特征，再经双向空间长短时记忆网络同步解析后形成长时空特征连接单元，并作为残差网络的输入。为了验证算法的有效性，使用 Kinect 传感器构建了一个全新的多模式手势数据集，在 3 个手势识别公开数据集 SLVM、Montalbano 和 SKIG 上的实验表明，提出的方法有很好的性能表现，识别精度超越了目前已公开的最佳识别率。

关键词：手势识别；3D 卷积神经网络；长段时记忆网络；宽残差网络

中图分类号：TP391.41　　doi: 10.3969/j.issn.1001-3695.2018.07.0429

Dynamic gesture recognition based on wide residual networks and long short-term memory networks

Liang Zhijie[1, 2], Liao Shengbin[1]

(1. *National Engineering Research Center for E-Learning*, *Central China Normal University*, *Wuhan* 430079, *China*; 2. *School of Adult & Network Education*, *Southwest University of Science &Technology*, *Mianyang Sichuan* 621010, *China*)

Abstract: The current dynamic hand gesture recognition method is not able to capture long-term spatiotemporal features from image sequences accurately. In order to solve this problem, this paper proposed a new dynamic gesture recognition algorithm based on wide residual networks and long short-term memory networks that perform simultaneous detection and classification. Firstly, spatial and temporal features are extracted from the fine-tuned 3D convolutional neural networks. Next, a bidirectional convolutional long short term memory networks is utilized to further take into account the temporal aspect of image sequences. Lastly, these higher level features are sent to the wide residual networks for final gesture recognition. In order to validate this method, a new challenging multimodal dynamic hand gesture dataset was introduced, which was captured with Kinect sensors. Experimental results show that proposed method achieves state-of -the-art performance on SLVM, Montalbano and SKIG.

Key words: gesture recognition; 3d convolutional neural networks; long short-term memory networks; wide residual networks

0　引言

手势是聋哑人群之间相互交流最有利的工具，也是聋哑人与正常人交流从而获取信息服务、共享社会物质文化成果最重要的途径。同时，手势具有自然、直观的视觉效果，因此在人机交互领域具有巨大的应用前景，越来越多的国内外相关研究机构和学者开始研究手势识别算法，以达到让机器自动理解人类手势的目标。然而，由于人手是复杂变形体，同时手势具有多样性、多义性，特别是动态手势还存在时间维度上的分布差异性，因此，当前手势识别仍然是存在诸多挑战的一项研究。

在人机交互领域最早的手势识别探索当属 Pavlovic 等人[1]利用人工神经网络以静止的手部形状为识别目标开展的研究。随着硬件系统的发展，静态手势识别中使用的算法复杂度也不断提高，Dardas 等人[2]采用支持向量机(support vector machine, SVM)算法研究了表示数字 0、1、…、9 的静态手势识别技术，并在实时环境中对识别方法进行了测试，取得了较高的识别率。通常来讲，静态手势识别只需要识别一张图片即可满足要求，而动态手势识别因其具有灵活多变、表意词汇丰富的特点，因此存在更多关于精度和可用性的困难。

当前大部分动态手势识别研究方法多是依靠人工经验进行特征提取：Parcheta 等人[3]基于隐马尔可夫模型(hidden Markov model, HMM) 对动态手势的识别进行了研究；在其自主创建的

收稿日期：2018-07-12；**修回日期**：2018-09-06　　**基金项目**：国家科技支撑计划项目（2015BAK3B02）；西南科技大学继续教育研究与发展基金资助项目（17JYF01）

作者简介：梁智杰（1987-），男，四川绵阳人，助理研究员，博士研究生，主要研究方向为计算机视觉和深度学习；廖盛斌（1969-），男（通信作者），副教授，博导，主要研究方向为大数据与机器智能、网络系统控制与优化（liaoshengbin@mail.ccnu.edu.cn）.

包含 91 个手语词汇的公开数据集上取得了 84.6%的识别率。Daniel 等人[4]提出了另外一种融合 HMM 和动态时间归整(dynamic time warping, DTW)的手语识别系统，该方法利用 HMM 进行人体上肢运动轨迹的准确跟踪，并在剑桥手势公开数据集 Cambridge Gestural Performance Database 2012 (CGPD12)[5]上取得了 95.1%的识别正确率。

在国内，曹洁等[6]首先利用 K-均值聚类算法对 RGB-D 图像完成动态手势的轮廓分割；然后结合快速动态时间规整算法完成动态手势识别，正确率达到了 96.8%。张备伟等[7]使用 Kinect 传感器获取人体关节点数据从而建立训练模板库；接着利用 DTW 算法完成了交警常用指挥手势的高精度识别。

然而，人工的特征提取和选择是一件非常耗时耗力的工作，必须要有非常深厚的专业知识和经验才能确保分类特征的正确性。同时，人工选取的特征也很难适应动态手势的多变性。近年来，随着计算机硬件性能的进一步提升，信息技术界迎来了又一轮人工智能变革的高潮，特别是基于神经网络的深度学习[8]受到了前所未有的关注。与传统人工特征提取加分类器的方式相比，深度学习方式将自动特征提取和分类联合为一体形成了“端到端”学习架构，避免了人工经验特征提取的主观性，因此在识别率上也取得质的提升。Moon 等人[9]研究了使用普通单目摄像头作为数据传感器的动态手势识别，提出了一种基于卷积神经网络(convolutional neural network, CNN)的大规模数据集手势识别方法。为了解决基于视频的动态手势识别需要对空间域和时间域进行同步特征提取这一问题，Molchanov 等人[10]首次提出了将仅能对图像进行特征提取的传统二维 CNN 模型扩展到可对空间和时间特征进行同步提取的三维模型，从而有效获取视频中的运动信息。

Wudi 等人[11]提出了一种双列深度网络的多模态手语识别方法：第一列 3D 卷积网络对视频数据进行运动特征的提取；第二列深度信念网络(Deep Belief Network, DBN)则利用骨骼数据进行识别。最后对两个子网络的分类结果进行有效的融合，从而在 Montalbano 手势识别大赛数据集上取得了 0.88 的 Jaccard Index 交并比得分[12]。Pigou 等人[23]则将 3D 卷积神经网络和循环神经网络(recursive neural network, RNN)两种模型进行叠加，在 Montalbano 数据集上将 Jaccard Index 交并比得分提升到了 0.916。

虽然当前基于深度网络架构的模型在手势的运动特征提取和分类上取得了较好的效果，但目前对视频中动态手语的识别仍然受限于长序列图像的处理。鉴于双向长短时记忆网络(bidirectional long short term memory, Bi-LSTM)在自然语言处理任务中表现出的优异性能，本文提出了一种全新的长序列手语识别深度学习架构 WRN-BCLSTM：首先以 3D 卷积神经网络作为视频的特征提取器，将其产生的固定长度短时空特征作为多层双向空间长短时记忆网络的输入，并进一步编码形成长时关联信息。而后通过宽残差网络(wide ResNet, WRN)对长序列视频的时空信息进行精确的表征。最后通过融合策略，对模型的两种独立特征分类结果进行有效的融合，以投票的方式获得视频中手势的类别。

概括来说，本文的创新点有：a)设计了适合短时特征提取的 3D 卷积层和适合长时信息编码的双向空间长短时记忆层，使模型能够最大程度上利用视频的时空特征进行分类。b)基于残差网络思想设计了用于特征选择的残差模块，拓宽了残差模块的卷积核宽度并减少了相应层数，从而增加了空间特征的选择范围，有效解决了深层网络的梯度衰减和梯度不均问题;c)提出了针对同源数据的有效融合策略，实现了在数据丢失时对单个分类器分类错误的补偿，使模型的分类准确率更高。

1　深度网络架构

针对动态手势识别的问题，提出了一种融合 3D 卷积神经网络、双向空间长短时记忆网络和宽残差模块的深度架构，如图 1 所示。

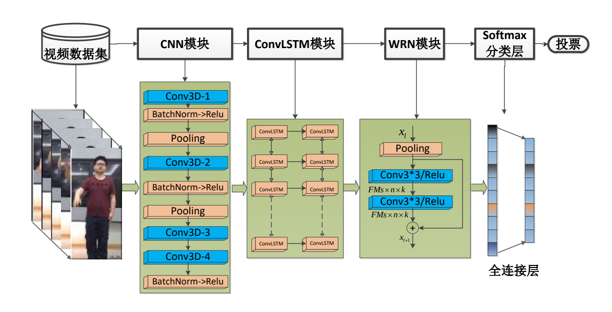


图 1　WRN-BCLSTM 模型结构图

首先，将包含手势样本的视频处理为具有统一长度的连续图像序列作为模型的输入。随后，利用 3D 卷积神经网络从图像序列中提取特征，空间维度和时间维度的特征数据经过双向空间 LSTM 解析后形成长时间动态特征数据连接单元，并以二维张量的形式作为残差网络的输入。经过残差层处理后最终被输入到一个 softmax 分类器，以向量的形式输出手势样本的预定义类别，而向量中每一个维度的值表示当前手势被分割到某个类别的置信度，即 $P(C|x,\theta)$ 。

1.1　数据预处理

深度网络架构由于其中全连接层的限制，一般都要求输入数据具有相同的维度。因此，首先需要对数据进行时间维度上的统一。为了尽可能精确地获取代表手势含义的特征，本文采用了窗口滑动法[12]，选择了 32 作为每一个手势视频的基准帧数。帧数大于 32 的视频，对两端无关图像序列进行删除，保留中间的关键帧，而对于帧数小于 32 的视频则是按照一定比例选出中间的若干帧进行插值。通过窗口滑动法的预处理，视频的运动路径信息得以保留。具体的预处理过程如下：

a)手势时间维度的分割。参照附图 2，使用窗口滑动法，将视频长度标准化为固定长度（譬如 32 帧）。如果采集的视频长度大于 32 帧，则删除两端的多余帧；反之，则重复某些帧。G_x^s 为原手语样例视频 x 的起始帧， G_x^e 为原手语样本视频 x 的结束帧， $L_x = G_x^e - G_x^s$ 为手语样例视频 x 的长度。

若 $L_x > 32$ ，则 $G_x^{ns} = G_x^s + (L_x - 32)/2$ ，这里 G_x^{ns} 为分割后的手语样例

x 新的起始帧。$G_x^{ne}=G_x^{ns}+32$，这里 G_x^{ne} 为分割后的手语样例 x 新的结束帧。

若 $L_x \le 32$，则 G_x^{s} 仍作为分割后的手语样例 x 的起始帧。$G_x^{ne}=G_x^{s}+32$，这里 G_x^{ne} 为分割后的手语样例 x 新的结束帧。这样，大部分带运动路径信息的关键帧得以保留。

b)手势的空间维度分割。按照人体区域范围将每一帧图像剪切为 112*112 像素，得到统一分辨率的视频。

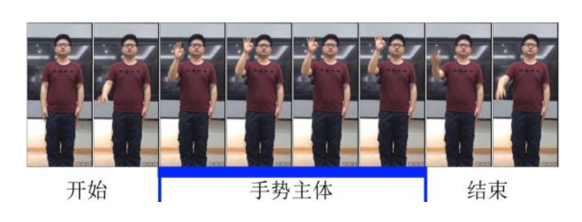


图 2　窗口滑动法

1.2　3D 卷积神经网络模块

卷积神经网络是一种特殊的前馈神经网络，它的三个重要特点，即局部连接、池化和权值共享使其非常适合图像数据的处理。局部连接保证了层级之间的稀疏性连接，大大降低了网络模型的参数规模。池化在一个小区域内采取一个特定的值作为输出，由此降低特征的维度。权值共享使同一感受视野内的神经元拥有相同的参数值，从而进一步简化网络结构，避免过拟合现象的发生。另外，由于卷积和池化的相互叠加决定了 CNN 具有一定程度上的平移、缩放和扭曲不变性[14]。

传统的 2D CNN 虽然对图像数据具有很强的特征提取能力，但是在面对视频任务处理时，由于时间维度被转换为长序列帧，因此容易丢失特征目标之间的运动信息。为了解决这一问题，本文采用了一种新的 3D CNN 结构对传统 2D CNN 进行改进。其中 3D 卷积定义如下：

$$f_{jk}^{xyz}=\sigma(b_j^l+\sum_m\sum_{p=0}^{P_{i-1}}\sum_{q=0}^{Q_{i-1}}\sum_{r=0}^{R_{i-1}}W_{ijm}^{pqr}v_{(l-1)m}^{(x+p)(y+q)(z+r)})$$

其中：f_{jk}^{xyz} 表示 3D 卷积操作的输出，$v_{(l-1)m}^{(x+p)(y+q)(z+r)}$ 表示输入样本视频数据的三个维度，上标中 x 和 y 分别代表输入样本的空间维度，z 代表输入样本的时间维度，p,q,r 分别表示本次卷积操作在三个维度上的值；下标中 $(l-1)m$ 表示第 $l-1$.层中的第 m 个特征图。W_{ijm}^{pqr} 是卷积核连接到前面第 m 个特征图中坐标为 (i,j,m) 的参数，也叫权值；P_i,Q_i,R_i 分别代表卷积核的尺寸；b_j^l 表示 l 层中的第 j 个特征图的偏置参数；$\sigma(\bullet)$ 是为了增强该结构的表达能力而引入连续的非线性激活函数。

由于传统的 sigmoid 和双正切 tanh 激活函数的导数值域都小于 1，梯度在经过每一层传递时都会不断衰减。因此，当网络结构不断加深时会出现梯度消失的问题。为了符合神经元的生物机理，本文使用了 rectified linear unit (Relu)作为激活函数，公式如下：

$$rectifier(X)=\max(0,X)$$

其中，当输入的 x 值小于等于 0 时，强制 x 等于 0；当输入的 x 值大于 0 时则不做改变。这样可以使输出具有一定的稀疏性从而加快网络的收敛速度。

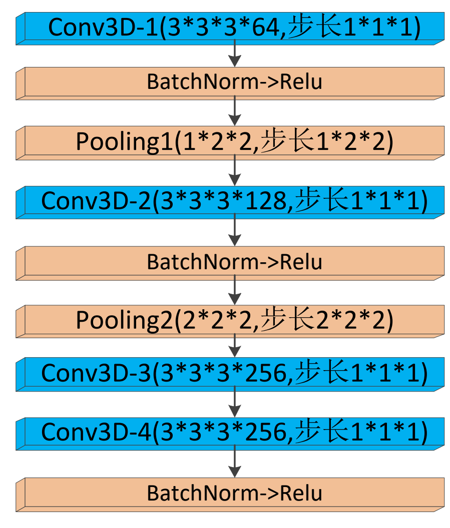


图 3　3D 卷积模块

本文使用的 3D 卷积模块结构如图 3 所示，网络的输入由连续的 32 帧图像构成，每一帧图像的空间尺寸为 112*112。3D 卷积 Conv3D-1 的卷积核尺寸为 3*3*3，特征个数为 64 个，每个 3D 卷积核具有相同的权重系数，输入数据经过卷积后得到 64 幅大小为 112*112*32 的特征图。同理，3D 卷积 Conv3D-2、Conv3D-3 和 Conv3D-4 层的卷积核个数分别为 128、256 和 256，尺寸统一保持为 3*3*3。池化层 Pooling1 只对空间维度进行 2*2 的降维采样，Pooling2 则从时间和空间维度同步进行 2*2*2 的采样，经过以上三次卷积和两次池化操作后得到 256 幅大小为 28*28*16 的特征图。在每一个卷积层之后，连接的是批规范化(Batch Normalization, BN)层，在梯度计算过程中对每个 mini-batch 的数据分布进行规范化，使其满足均值为 0，方差为 1，再输入到下一层计算。通过使用批规范化，保证初始学习速率可以选择相对较大的值，以提高收敛速度。

1.3　双向空间 LSTM 网络模块

动态手势识别的目标是从视频序列中提取出手势的时空视觉信息，但视频事件的时序往往比较复杂，这给识别任务带了挑战。鉴于 LSTM 近年来在自然语言处理领域处理复杂时序任务时取得了巨大成功，本文探索使用 LSTM 从输入的视频中递归学习出图像序列的长时间动态特征。

作为循环神经网络的一个变体，LSTM 网络依靠记忆单元 c_t 来记录序列到当前时刻为止所有的历史信息，并使用输入门 i_t，遗忘门 f_t 和输出门 o_t 来控制梯度在时间维度上依次传播，进而能够将输入的序列 (x_1,L,x_t) 映射为隐藏结点序列 (h_1,L,h_t)，从而可以从输入序列的动态特征中递归学习到复杂的时间关联信息。然而，传统的自然语言处理领域中应用的 LSTM 是将一维的向量作为处理对象，主要学习一段文字向量化后的时序特征，如果将这种结构直接应用到视频分类任务中，不可避免的存在图像空间位置信息的丢失。2015 年 NIPS 会议，Shi 等[15]提出了 Convolutional LSTM (ConvLSTM)，该模型可以直接对二维张量进行运算，有效克服了时序传递过程中空间信息丢失这一问题，并在视频事件分析领域获得成功。以此结构为基础，我们设计了双向空间 LSTM，其结构如图 4 所示

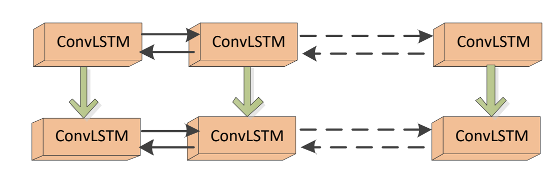


图 4　双向空间 LSTM 模块

两个单向传递的 ConvLSTM 连接后构成双向(Bi-ConvLSTM)记忆单元。其中每一个 Bi-ConvLSTM 记忆单元包含了来自 3D 卷积模块的空间和时间的输入，Bi-ConvLSTM 单元的计算结构可以描述为

$$i_t = s(U_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \bullet c_{t-1} + b_i)$$

$$f_t = s(U_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \bullet c_{t-1} + b_f)$$

$$o_t = s(U_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \bullet c_t + b_o)$$

$$c_t = f_t \bullet c_{t-1} + i_t \bullet y(U_{xc} * x_t + W_{hc} * h_{t-1} + b_c)$$

$$h_t = o_t \bullet y(c_t)$$

其中：x_t 表示当前时刻的输入，h_{t-1} 表示 $t-1$ 时刻的输出，且都是以二维张量的形式存储。$*$ 表示为卷积操作，$\bullet$ 表示为哈达玛积(Hadamard product)。i_t、f_t 和 o_t 分别表示为输入门、遗忘门和输出门，U、W 和 b 分别表示上述三种门结构的输入权重、递归权重和偏置项，i_t 决定了一个内存单元加入多少新的信息，f_t 控制每一个内存单元需要遗忘掉多少信息，o_t 控制每一个内存单元输出多少信息。$s(x)=(1+e^{-x})^{-1}$ 表示 sigmoid 非线性函数，使得三个门的元素取值在[0,1]之间，$y(x)=(e^x-e^{-x})/(e^x+e^{-x})$ 表示双曲正切非线性函数，取值范围是[-1,1]，c_t 表示为输入控制模块，是 ConvLSTM 的核心记忆单元，控制了哪些信息将被保存；c_t 由两部分组成，第一部分表示上一时刻记忆单元 c_{t-1} 经过遗忘门 f_t 后留下的信息，第二部分是输入数据经过调制门后留下的信息。

图像序列的时空信息经过双向空间 LSTM 的传递后，在全局范围内得到了有效的融合，相比单向的 LSTM，双向 LSTM 网络能够更好地捕捉到视频的全局信息，因此能够获得更好的预测结果。

1.4　宽残差网络结构设计

经过 3DCNN 和 Bi-ConvLSTM 叠加结构的编码，手势视频被转换成了蕴含大量时空信息的二维张量特征。近年来崛起的深度卷积神经网络在图像分类任务中表现出了的优异的性能。特别是 2015 年之后，AlexNet，GoogleNet，VGG，ResNet 等网络模型的进展使得深度学习架构在图像分类任务中取得了连续性的突破。基于此，本文结合当前性能较为出色的深度残差架构思想进行创新尝试，对二维图像特征进行高效准确的自动学习和分类。

理论上来说，模型架构的容量和特征判别能力能够随着网络层数的不断加深而不断提高。然而大量实践尝试结果表明，简单增加网络的深度会出现梯度弥散问题，即过深的网络结构易导致训练无法收敛，因此识别率反而降低。针对该问题，何凯明等人[16]提出了使用捷径连接（shortcut connection，SC）搭建深度残差网络结构。

假设 $H(x)$ 表示深度神经网络在输入样本 x 后的最优解映射，传统的卷积神经网络是直接拟合 $H(x)=x$，而深度残差网络期望拟合残差映射，即 $F(x)=H(x)-x$。由于 x 是输入的源图像，可以验证拟合 $F(x)$ 与拟合 $H(x)$ 的目标是等价的。在此条件下，原来的最优解映射被表示为：$H(x)=F(x)+x$。如图 5(a)所示。

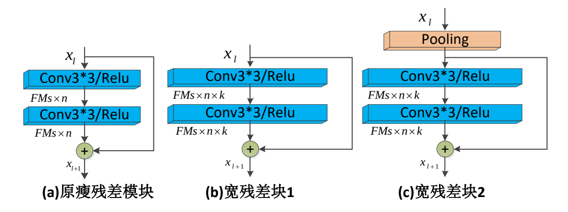


图 5　宽残差模块示意图

此时，深度残差网络通过快捷连接结构，跳过 2 或 3 个卷积层，自身映射到叠加层与卷积层的输出进行相加。显然，使网络去拟合确定的函数 $F(x)=0$ 比优化逼近一个最优函数 $H(x)$ 要容易很多。在快捷连结方式中，模型的参数量和计算复杂度并没有大的增加。换而言之，深度残差结构可以叠加到已有的深度模型中去，而不改变原模型已有的架构，使得训练出性能更好、层数更多的网络模型成为可能。

但即使如此，具有恒等映射的残差网络随着深度的不断增加也同样存在着多层残差模块共享少量的梯度信息流这一弊端。换而言之，只有少部分残差模块的参数得到了更新。为了解决该问题，本文结合 Sergey 等人[27]提出的宽度残差模块思想，使用浅而宽的结构代替了深而窄的残差网络模块。后续实验证明，适当增加残差模块的宽度比增加单纯增加网络的深度更能提高残差网络的性能，因为更宽的网络增加了特征的选择范围，从而增强了特征的耦合能力。

如图 5 所示，本文的宽残差(Wide Residual Network, WRN)模块从 Bi-ConvLSTM 输出的二维张量中进一步提取数据的空间特征。宽残差模块的总层数为 16，一共由 4 个残差组 (Conv1、Conv2、Conv3 和 Conv4) 构成，残差组的宽度由加宽系数 k 决定，本文中 $k=4$。每一个残差组中又包含了 $N=4$ 个残差块。第一和第二个残差组 Conv1、Conv2 中的宽残差块对应图 5(b)，第三和第四个残差组 Conv3、Conv4 对应图 5(c)，也就是在第一层进行空间的池化。这样以来，每一层的特征图个数分别是 8*4、16*4、32*4 和 64*4，从而在有效降低了残差模块层数的同时拓宽了卷积核的个数，而模型的参数量并没有因此而增加。

2　模型优化

模型的优化可以理解为通过训练样本和验证样本对模型的性能进行初步测评，并选择合适的超参数训练出一个最优的决策模型用于最终的测试样本。

2.1　损失函数和正则化

网络架构的优化通过计算损失函数来实现，本文的输出层使用的是 softmax，分类层的输出按以下公式计算：

$$P(C|x,\theta) = soft\max(x_k) = \frac{\exp(x_k)}{\sum_{i=1}^{K}\exp(x_i)}$$

这里 x_i 代表第 i 个神经元的输出。θ 代表模型参数，P 代表网络输出层对应某个手势的概率。考虑到多分类的计算问题，损失函数使用了负对数似然函数来反映网络输出和实际手势标签之间的差异性：

$$L(\theta,D)=-\frac{1}{|D|}\sum_{i=0}^{|D|}\log(P(C^{(i)}|x^{(i)},\theta))$$

其中：x 为输入样本对应的特征表示，C 表示目标类别标签，θ 为待优化的模型参数，D 表示同一批输入样本的数量(mini-batch size)，网络的优化就是一个通过修改参数 θ 来不断减少误差 L 的过程。

为了解决过拟合问题，本文在经验风险最小化原则上给 softmax 的 Loss 加上了正则化项：

$$L=L_0+\lambda\|\theta\|^2$$

其中：第一部分 L_0 对应原损失函数，第二部分 $\lambda\|\theta\|^2$ 是 L2 范数的正则化项，用来减少参数的优化空间，从而避免过拟合。λ 是正则化系数，用来控制该正则化对损失函数所起到的约束强度，λ 的值可以通过交叉验证(cross validation)来选择。

2.2　参数优化

参数优化是指由损失函数计算得到的误差来反向传播从而计算每一层参数的梯度，本文使用了一种改进的梯度下降优化算法进行神经网络参数的更新：

$$\nabla f(\theta_t)=\langle\frac{\delta L}{\delta(\theta_{t-1})}\rangle_{batch}$$

$$v_{t+1}=\mu v_t-\varepsilon\nabla f(\theta_t+\mu v_t)$$

$$\theta_{t+1}=\theta_t+v_{t+1}$$

其中：$\nabla f(\theta_t)$ 表示使用一个批量（batch）的数据训练后得到的损失函数 L 相对于前一个迭代周期参数 θ_{t-1} 的梯度，且第 t 次迭代时的参数更新依赖于发生在第 $t-1$ 次迭代时的更新。ε 表示学习速率，因为网络结构中使用了批规范化，因此可以为学习速率 ε 设定一个稍微大的初始值；为了防止过拟合，在迭代过程中如果损失函数误差值的减小速率没有达到预期，则进行相应的权值衰减，从而保证参数更新幅度不断减小，使学习过程向着复杂决策面的反方向偏置。v_t 是动量项表示当前迭代累计的参数调整惯性。μ 是冲量系数设为 0.9，在迭代的初期，使用前一次的梯度进行加速；而在迭代后期优化到达收敛值附近时，因为两次更新方向基本相反，使得梯度逐渐缩小。这里使用的参数更新法则与随机梯度下降 (stochastic gradient descent, SGD)的方法类似，不同点是这里在计算梯度的时候，求解权重加上了冲量(momentum)的梯度 $\nabla f(\theta_t+\mu v_t)$，而 SGD 中只是简单的计算当前权重的梯度 $\nabla f(\theta_t)$，所以收敛速度相比传统 SGD 有了很大的提升。

2.3　多模式融合

在训练样本有限的情况下，特征融合被证明是进一步提升识别效果的有效手段[17]。如结构图 1 所示，本文使用了一种双列深度结构对输入的视频提取不同的特征，每一个子网络以不同的数据格式作为输入，因此识别效果也不尽相同。测试阶段的融合模型按照下式结合从两个子网络估计出的所属类别概率来计算手语分类的最终输出：

$$P(C|x)\propto a*P(C|x_I,W_I)+(1-a)*P(C|x_c,W_C)$$

此处，不同子网络输出的所属成员概率由 softmax 函数得出，a 代表一个加权系数，在训练阶段由交叉验证得出，用来控制每一个子网络对最终的成员概率 P 的贡献。一般来说，a 的值非常接近 0.5。本文根据不同的数据输入格式，分别训练两个 WRN-BCLSTM 模型并将输出结果进行概率融合，从而达到鲁棒性强，实时性高、正确率高的目的。

3　实验结果与分析

本文的实验环境如下，操作系统：64 位 Ubuntu16.04 LTS；CPU：Inter Core i7-6700K 八核；显卡 Nivida GeForce GTX1070 11264M 显存；32 GB DDR4 内存；实验框架选择了 Tensorflow。数据集选用了博物馆聋哑人手语数据集（sign language video in museums, SLVM）、ChaLearn Looking at People 2014 Gesture datasets(简称 Montalbano)以及 Sheffield Kinect Gesture (SKIG) 公开数据集。

3.1　SLVM 数据集的实验结果

数据是进行手势识别研究的重要基础和先决条件。然而，当前大多数的公共数据集缺乏有效和准确的标签，或以单一的数据格式进行存储。为了满足长序列动态手语识别研究的需求，本文设计了多模式同源信号的数据采集平台，并建立了聋哑人在博物馆参观过程中使用的高频手语词汇数据样本集。

首先，在数据采集模块，为了有效抑制光照和场景噪声的干扰，本文摒弃了以往传统的使用 RGB 图像作为训练样例的方法，而是基于 Kinect V2 for Windows 开发了多模态数据采集系统 Gestures Recorder(http://pan.baidu.com/s/1dEX29R7)。

如图 6 所示，该系统从红外图像、轮廓图像、骨骼数据中同步进行特征保存，采集动态手语词汇 20 类，共计 6800 个样本，其中训练样本(training data) 5100 个，验证样本(validation data) 850 个，测试样本(test data) 850 个，视频分辨率为 512*424，形成了一套完整的手语数据库(https://pan.baidu.com/s/1pL2qwuZ)。

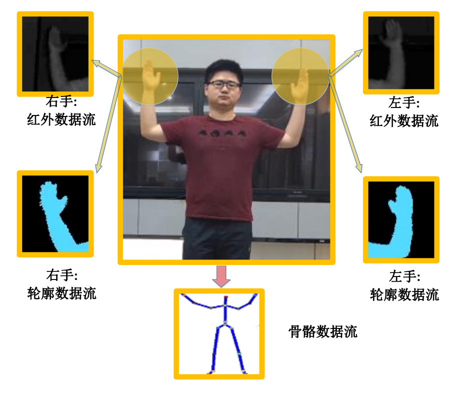


图 6　SLVM 多模式同源数据集

本文使用了迁移学习(transfer learning)的思想来缩短模型训练时间。迁移学习是指将原任务领域学习到的参数信息共享

和推广到相似的学习任务中以提高模型的识别率。以动态手势识别为例，如果完全从初始状态训练深度网络架构，可能会因为数据样本不足而无法达到预期的效果。为缩短训练时间，本文采用了两种迁移学习策略：第一步是模块迁移，在文献[13]的网络基础上将该模型在大规模视频数据集 IsoGD (Chalearn LAP 2017 RGB-D isolated gesture dataset)上已经训练好的模型前 9 层 3DCNN 模块迁移到现有模型中与 Bi-ConvLSTM 和 WRN 模块进行叠加。由于模型的 3DCNN 层提取的是图像序列的边缘、色彩以及短时空特征等信息，在视频分类领域具有一定的共性。所以，在参数迁移过程中可以固定前 9 层的参数并进行调整和优化。第二步是数据迁移，根据确定的 3DCNN+Bi-ConvLSTM+WRN 的网络架构，在 IOSGD 大规模数据集上进行分类学习得到网络预训练模型，而后将模型迁移至 SLVM 数据集上替换掉分类输出层，并进行参数微调。学习速率的设置采用了均匀分布策略，以 0.05 作为初始的学习速率，经过 1920 次迭代后乘以 0.1，batch size 设置为 8。实验在 GPU 加速基础上，可以在 2 小时完成一个 epoch 的迭代，经过 12 个 epoch 的迭代，网络已经收敛的非常好。

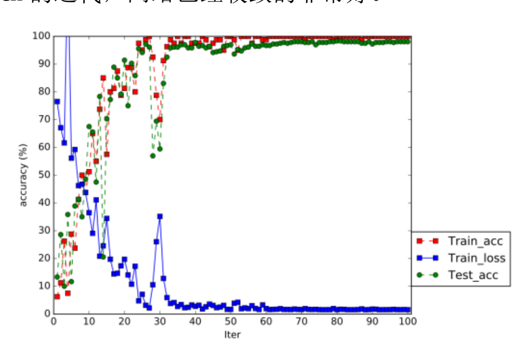


图 7　网络在 SLVM 数据集的迭代过程

图 7 是网络的识别正确率曲线，横坐标为训练的迭代次数，在 SLVM 数据集上的打印频率设置为 76 个 iter 输出一个阶段结果，纵坐标表示正确率。最佳的识别率达到了 98.3%。

表 1　算法在 SLVM 数据集上的方法比较

模型	数据类型	识别率
3DCNN[21]	轮廓图+骨骼	80.8%
	+Relus	85.5%
	+L2 正则化	87.6%
	红外图+骨骼	81.5%
	+Relus	86.1%
	+L2 正则化	88.3%
	数据融合	89.2%
本文方法	轮廓图	90.1%
	+Relus	95.2%
	+L2 正则化	97.7%
	红外图	90.3%
	+Relus	95.3%
	+L2 正则化	98.1%
	数据融合	98.3%

表 1 将本文的方法和在 SLVM 数据集上已公开的最佳方法进行了对比，实验结果表明，本文的方法相比[21]的模型具有极大的优势：首先，3DCNN 通过卷积操作共享网络层参数，在有效降低了网络参数量的同时，有效地对图像序列的局部、甚至是整体的空间特征信息进行有效提取。Bi-ConvLSTM 不但具有传统长短时记忆网络的自适应记忆和抗遗忘的能力，在进行时间序列学习时更加关注图像序列的空间信息。因此本文提出的模型相比之前 SLVM 数据集的最佳识别正确率模型有质的提升。同时可以发现，当选择使用 Relu 激活函数时，网络具有更强的泛化能力，而使用 L2 正则化也可以在一定程度上避免过拟合现象。

3.2　Montalbano 手势数据集的实验结果

为了验证本文算法在大规模数据集上的有效性，本文选用了 Montalbano 数据集进行实验对比分析。该手语数据集是用深度摄像机录制的多人动态手语数据集，旨在实现基于多模态数据的非特定用户动态手势识别。本手势识别大赛在 2014 年举办，包含了意大利语中的 20 个常用手势表达。数据集包含了 13858 个数据样本（其中 training data 7754 个，validation data 3362 个，test data2742 个），每一个多模态样本包括了传统的 RGB 图像，深度图像，骨骼数据和轮廓图像。该公共数据的详细情况可见文献[12]。

本文按照 Montalbano 手语大赛的规则，以 Jaccard Index 得分对算法的性能进行综合测评。其中大赛官方的 Jaccard Index 测评方法如下：

$$J_{(s,n)} = \frac{\left|A_{(s,n)} \cap B_{(s,n)}\right|}{\left|A_{(s,n)} \cup B_{(s,n)}\right|}$$

对于一个动态手势，有其所属类别的真实边界 A 和算法预测识别输出边界 B， J 表示两个边界所占区域的交集与其并集的比值。

$$J_{(mean)} = \frac{1}{NS}\sum_{s=1}^{S}\sum_{n=1}^{N} J_{(s,n)}$$

其中： N 表示了数据集中的手势类别个数，此处 $N=20$； S 表示被测试样例中的帧序列长度。 $J_{(mean)}$ 表示对测试数据中所有样本的 Jaccard Index 交并比例取均值作为算法在该数据集上的最终得分。

考虑到训练一个复杂的深度网络是一件非常耗时的工作，特别是在 Montalbano 这种大型的数据集上，因此本文将之前在 IsoGD 数据集上训练完成的网络架构作为初始化模型，而后在 Montalbano 数据集上调参。以 0.01 作为初始的学习速率，每经 2500 次迭代后衰减至 1/10，batch size 设置为 8，通过实验可以发现，在数据样本有限的情况下，迁移预训练模型相比从原始状态开始训练不仅节省了训练时间，同时使识别正确率有了较大的提升。

如图 8 所示，在 Montalbano 数据集上的打印频率为 95 个 iter 显示出一个阶段结果。由于迁移了预训练模型，网络输出函数的损失值(loss)下降幅度很快，这也为网络进一步优化节省

了时间。

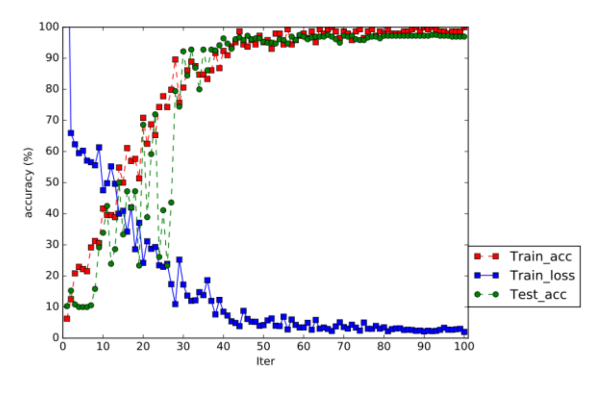


图 8　网络在 Montalbano 数据集的迭代过程

表 2 对各种算法的结果进行了对比。其中前三项工作[18~20]使用的是基于手工特征提取的方法，可以发现，深度学习在有效特征的自动提取方面相比手工特征提取具有巨大的优势。文献[21]方法是本文的前期方法，采用了 3DCNN 结合数据融合技术进行手势识别，证明了有效的数据融合技术相比单个网络模型可以将整体识别率提升了近 5%。

文献[11]方法利用基于受限玻尔兹曼机(restricted Boltzmann machines, RBM)的概率型自编码器来处理骨骼数据，利用 CNN 处理 RGB-D 图像。同时，在两个并行的子网络顶端利用隐马尔可夫模型有效地将静态分类器的优势发挥到处理动态序列模式任务上，取得了较好的效果。方法[22]提出了用神经网络学习音频+视频模态特征的应用，证明了多模态融合会比单个模态学到更好的特征，此方法在该届手势识别大赛中取得了第一名。

文献[23]方法借鉴了 RNN 模型在自然语言处理任务中所取得的巨大成功，首次将 CNN+RNN 的模型应用到了手势识别领域，使该数据集的正确率取得了质的提升，证明了 RNN 模型在处理序列样本时的优异性能。与方法[23]相同，本文也是采用 3D 卷积神经网络自动的从视频中进行短时特征提取。不同之处在于，本文结合空间卷积 LSTM 递归网络，在对长序列特征进行学习的同时，不丢弃视频样本序列之间的空间关联信息，给分类带来了更为精确的结果。其中，使用深度视频的识别结果略高于 RGB 视频，是因为 RGB 图像容易受环境光照和复杂背景的影响。而使用融合策略的结果要优于任何一种单一特征。证明了本文提出的特征融合方法是行之有效的。

3.3　SKIG 手势数据集的实验结果

为了更进一步验证算法对动态手势分类的有效性，本文选用了 Sheffield Kinect Gesture Dataset(SKIG)进行实验对比分析。该数据集由 Ling Shao 等[24]建立，包含了 2160 个独立的手势样本，由 1080 个 RGB 和 1080 个 Depth 格式的视频保存。每个样本包含 1 个独立的动态手势，共计 10 个类别：circle, triangle, up-down, right-left, wave, Z, cross, come here, turn around 和 Pat。为了体现数据在真实环境下的复杂背景和光照多样性，该数据集由 6 个手势执行者分别在三种背景（白色、木纹和有字的报纸）和两种光照（强光、弱光）条件下使用三种不同的姿态（握拳、单个手指和手掌）来录制。这些样本虽然在人眼看来是同一个类别，但对于网络模型而言，不同的位置、背景和光照条件代表着不同的神经元响应。因此，本数据集对验证算法在不同模式以及执行环境差别较大时的辨识能力很有价值。

表 2　算法在 Montalbano 数据集上的方法比较

模型	数据类型	Jaccard
Random Forest[18]	RGB 视频	0.787
MRF[19]	骨骼+RGB	0.826
Boosted classifier[20]	Skeleton+Depth+RGB	0.833
	RGB+骨骼	0.817
3DCNN[21]	Depth+骨骼	0.829
	数据融合	0.836
	Skeleton	0.863
Dynamic DNN [11]	RGB-D	0.787
	数据融合	0.879
DNN[22]	Depth+RGB+Audio	0.881
RNN+3DCNN[23]	Depth+RGB+skeleton	0.916
	RGB	0.921
本文的方法	Depth	0.926
	数据融合	0.932

在该公开数据集上的实验设计采用了与文献[24]相同的原则：3 折交叉验证，即每一次用 4 个执行者的样本分别做训练和验证集合，其余 2 个执行者的样本用于最终的测试集。为了利用不同学习任务之间的共性来缩短模型的训练时间，实验以 Montalbano 数据集上训练完成的网络为初始化模型，将输出层节点个数改为适应 SKIG 数据集的参数，分别以微调的方式将两种数据集的 RGB 视频和 Depth 视频进行对接来训练两个子网络。以 0.01 作为初始的学习速率，每经 150 次迭代后衰减至 1/10。

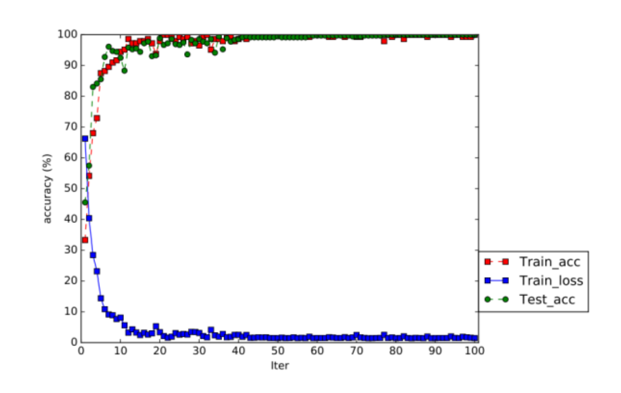


图 9　网络在 SKIG 数据集的迭代过程

如图 9 所示，batch size 值取 8，网络训练了 10 个周期，每 9 个 iter 打印一次结果。由于采用了迁移学习策略，在训练过程的早期，模型就展现出了较强的学习能力，损失函数值下降很快，在迭代 4 个周期次后，损失函数的值趋于稳定，经历了 8 个周期之后，损失函数值接近与 0，此时，网络训练过程已经接近收敛。从识别正确率显现出了类似的趋势：test 正确率在识别早期上升较快，迭代后期趋于稳定，达到了 99.3%的识别

率。

表3列举了各种公开方法在SKIG测试集上取得的正确率，本文方法在SKIG数据集上同样取得了较好的识别结果。

表3　算法在SKIG数据集上的方法比较

模型	识别率
RGGP+RGB-D[24]	88.7%
DLEH2(DLE+HOG2)[26]	98.4%
3DCNN+RNN+CTC[25]	98.3%
3DCNN+ConvLSTM[13]	98.9%
本文方法	99.3%

其中，作为该公共数据集的基准方法，文献[24]提出了一种融合RGB和深度数据的自适应方法，并采用了基于图的受限制遗传算法对模型进行优化，从而对动态手势进行了有效的分类。文献[26]使用了一种基于形状和时空特征深度融合的动态手势识别方法，将两种数据的特征描述分别与线性SVM分类器结合，将SKIG的识别率提升到了98.4%。文献[25]首次将卷积神经网络和循环神经网络叠加用于该数据集：首先利用3D卷积神经网络提取出手势的运动的三维时空信息，并根据这些信息使用循环神经网络捕捉到长序列的时空特征，进而实现了高精度的动态手势分类。文献[13]首次将空间LSTM模型用于手势分类，并在分类器之前使用了金字塔池化（spatial pyramid pooling，SPP）从不同尺寸的特征图中进行空间信息提取和聚合，有效节省了训练时间。本文使用双向空间LSTM最大程度上利用视频的空间信息融合长序列信息，并使用宽残差网络对视频转换成的二维张量进行准确分类，因此识别率也得到了进一步的提高。

4　结束语

本文设计了一种融合宽度残差模块和双向空间长短时记忆网络的深度模型WRN-BCLSTM，并用于动态手势识别任务。该模型以3D卷积神经网络从视频流中提取空间特征和短时动态特征，以双向空间长短时记忆网络从3DCNN的输出中充分获取特征的上下文长时信息，并在模型的后半段引入宽残差模块，较好的解决了传统深度堆叠残差模块的特征衰减和重用问题，提高了模型了辨识能力。

在三个公开的动态手势数据集上的实验验证了模型的有效性。从实验结果可以发现：(1)将两种数据流分别输入到网络，并将输出的结果进行融合是行之有效的。相比任何一种单一数据流特征，融合策略对识别性能的提升是显著的。(2)ConvLSTM不仅具有传统LSTM的时序建模能力，而且还可以像CNN一样刻画空间局部特征，因此更加适用于空间时序信息的处理。(3)适当增加宽残差模块的宽度比单一增加残差网络的深度更能提高残差网络的性能，且宽残差网络的训练速度更快。(4)在数据样本有限的情况下，采用迁移学习策略不仅有利于缩短训练时间，而且使网络具有更好的泛化效果。

深度学习技术从原始的数据空间逐层进行特征抽象和变化，最终利用高层特征进行分类，极大地提升了动态手势识别的准确性。然而，深度模型仍然存在着很多未知因素可以探索。比如，如何设计适用于3D卷积模块的残差快捷连接方式等。同时，还有很多其他的深度学习网络结构，比如基于注意力模型(attention model)的深度网络在人体动作理解领域展现出了很大的优势，这些将是我们未来的研究方向。

参考文献：

[1] Sharma R, Pavlovic V I, Huang T S. Toward multimodal human-computer interface [J]. Proceedings of the IEEE, 1998, 86 (5): 853-869.

[2] Dardas N H, Georganas N D. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques [J]. IEEE Trans on Instrumentation & Measurement, 2011, 60 (11): 3592-3607.

[3] Parcheta Z, Martínez-Hinarejos C D. Sign language gesture recognition using hmm [C]// Proc of Iberian Conference on Pattern Recognition and Image Analysis. Berlin: Springer, 2017: 419-426.

[4] Radu-Daniel V. Beyond features for recognition: human-readable measures to understand users'whole-body gesture performance [J]. International Journal of Human-Computer Interaction, 2017, 12 (23): 1-16.

[5] Simon F, Helena M, Pushmeet K, *et al.* Instructing people for training gestural interactive systems [C]// Proc of SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2012: 1737-1746.

[6] 曹洁，赵修龙，王进花. 基于RGB-D信息的动态手势识别方法 [J]. 计算机应用研究, 2018, 35 (7): 2228-2232. (Cao Jie, Zhao Xiulong, Wang Jinhua. Dynamic gesture recognition approach based on RGB-D information [J]. Application Research of Computers, 2018, 35 (7): 2228-2232.)

[7] 张备伟，吴琦，刘光徽. 基于DTW的交警指挥手势识别方法 [J]. 计算机应用研究, 2017, 34 (11): 3494-3499. (Zhang Beiwei, Wu Qi, Liu Guanghui. Method for recognizing gesture of traffic police based on DTW algorithm [J]. Application Research of Computers, 2017, 34 (11): 3494-3499.)

[8] Mohanty Aparna, Sahay Rajiv R. Understanding Indian classical dance by recognizing emotions using deep learning [J]. Pattern Recognition, 2018, 7 (79): 97-113.

[9] Hyeon Chul Moon, Anna Yang, Jae-Gon Kim. CNN-Based Hand Gesture Recognition for Wearable Applications [J]. Journal of Broadcast Engineering, 2018, 3 (23) 246-252.

[10] Molchanov Pavlo, Gupta Shalini, Kim Kihwan, *et al.* Hand gesture recognition with 3D convolutional neural networks [C]// Computer Vision and Pattern Recognition Workshops. New York: IEEE, 2015: 1-7.

[11] Wu Di, Lionel Pigou, Pieter-jan Kindermans, *et al.* Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2016, 38 (8): 1583-1597.

[12] Wan Jun, Sergio Escalera, Gholamreza Anbarjafari, *et al.* Results and Analysis of ChaLearn LAP Multi-modal Isolated and Continuous Gesture Recognition, and Real Versus Fake Expressed Emotions Challenges [C]// IEEE International Conference on Computer Vision Workshops. New York: IEEE, 2018: 3189-3197.

[13] Zhu Guangming, Zhang Liang, Shen Peiyi, *et al.* Multimodal Gesture Recognition Using 3D Convolution and Convolutional LSTM [J]. IEEE Access, 2017, 5 (99): 4517-4524.

[14] Tu Zhigang, Xie Wei, Qin Qianqing, *et al.* Multi-stream CNN: Learning representations based on human-related regions for action recognition [J]. Pattern Recognition, 2018, 2 (79): 32-43.

[15] Shi Xinjian, Chen Zhourong, Wang Hao, *et al.* Convolutional LSTM Network: a machine learning approach for precipitation now casting [C]// Proc of International Conference on Neural Information Processing Systems, Massachusetts: MIT Press, 2015: 802-810.

[16] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep Residual Learning for Image Recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Computer Society, 2016: 770-778.

[17] Guo Dan, Zhou Wengang, Li Houqiang, *et al.* Online Early-Late Fusion Based on Adaptive HMM for Sign Language Recognition [J]. ACM Trans on Multimedia Computing Communications & Applications, 2018, 14 (1): 1-18.

[18] Necati Cihan, Ahmet Alp kindiroglu, Lale Akarun. Gesture Recognition Using Template Based Random Forest Classifiers [C]// Proc of Computer Vision. Zurich: Springer International Publishing, 2014: 579-594.

[19] Chang, Juyong. Nonparametric Gesture Labeling from Multi-modal Data [C]// Proc of Computer Vision Workshops. Zurich: Springer International Publishing, 2014: 503–517.

[20] Camille Monnier, Stan German, Andrey Ost. A Multi-scale Boosted Detector for Efficient and Robust Gesture Recognition [C]// Proc of Computer Vision Workshops. Zurich: Springer International Publishing, 2014: 491–502.

[21] Liang Zhijie, Liao Shengbin, Hu Bingzhang. 3D convolutional neural networks for dynamic sign language recognition [J]. Computer Journal, 2018, 5 (4): 1-13.

[22] Neverova Natalia, Wolf Christian, Taylor Graham, *et al.* ModDrop: Adaptive Multi-Modal Gesture Recognition [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2016, 38 (8): 1692-1706.

[23] Pigou Lionel, Van Den Oord Aaron, Dieleman Sander, *et al.* Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video [J]. International Journal of Computer Vision, 2015, 5 (3): 1-10.

[24] Li Liu, Ling Shao. Learning discriminative representations from RGB-D video data [C]// Proc of International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2013: 1493-1500.

[25] Pavlo M, Yang Xiaodong, Gupta Shalini, *et al.* Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks [C]// Computer Vision and Pattern Recognition. New York: IEEE, 2016: 4207-4215.

[26] Zheng Jinqing, Feng Zhiyong, Xu Chao, *et al.* Fusing shape and spatio-temporal features for depth-based dynamic hand gesture recognition [J]. Multimedia Tools & Applications, 2017, 5 (76): 1-20.

[27] Zagoruyko S, Komodakis N. Deep compare: a study on using convolutional neural networks to compare image patches [J]. Computer Vision & Image Understanding, 2017, 16 (64): 1-9.